Affect Sensing and Contextual Affect Modeling from Improvisational Interaction

Li Zhang l.zhang@tees.ac.uk
School of Computing
Teesside University
Middlesbrough, TS1 3BA, UK

Abstract

We report work on adding an improvisational AI actor to an existing virtual improvisational environment, a text-based software system for dramatic improvisation in simple virtual scenarios, for use primarily in learning contexts. The improvisational AI actor has an affect-detection component, which is aimed at detecting affective aspects (concerning emotions, moods, value judgments, etc.) of human-controlled characters' textual "speeches". The AI actor will also make an appropriate response based on this affective understanding, which intends to stimulate the improvisation. The work also accompanies basic research into how affect is conveyed linguistically. A distinctive feature of the project is a focus on the metaphorical ways in which affect is conveyed. Moreover, we have also introduced affect detection using context profiles. Finally, we have reported user testing conducted for the improvisational AI actor and evaluation results of the affect detection component. Our work contributes to the journal themes on affective user interfaces, affect sensing and improvisational or dramatic natural language interaction.

Keywords: Affect detection, metaphorical language, intelligent conversational agents, dramatic improvisation and context profiles.

1. INTRODUCTION

In our previous work, we have developed online multi-user role-play software that could be used for education or entertainment. In this software young people could interact online in a 3D virtual drama stage with others under the guidance of a human director. In one session, up to five virtual characters are controlled on a virtual stage by human users ("actors"), with characters' (textual) "speeches" typed by the actors operating the characters. A graphical interface on each actor's and director's terminal shows the stage and characters. Speeches are shown as text bubbles. Actors and the human director work through software clients connecting with the server. The actors are given a loose scenario around which to improvise, but are at liberty to be creative.

The human director needs to constantly monitor the unfolding drama and the actors' interactions, or lack of them, in order to check whether they are keeping to the general spirit of the scenario. If this is not happening, the director may then intervene. Director's interventions may take a number of forms. Director may choose to send messages to actors or may introduce and control a bit-part character. This character may not have a major role in the drama, but can help to stimulate the improvisation. But this places a heavy burden on directors, especially if they are, for example, teachers and unpracticed in the directorial role.

One research aim is thus partially to automate the directorial functions, which importantly involve affect detection. For instance, a director may intervene when emotions expressed or discussed by characters are not as expected. Hence we have developed an affect-detection module. The module identifies affect in characters' text input, and makes appropriate responses to help stimulate the improvisation. Within affect we include: basic and complex emotions such as anger and embarrassment; meta-emotions such as desiring to overcome anxiety; moods such as hostility; and value judgments (of goodness, etc.). Although merely detecting affect is limited compared to extracting full meaning, this is often enough for stimulating improvisation. The results of this affective analysis are then used to: (a) control an automated improvisational AI actor – EMMA (emotion, metaphor and affect) that operates a bit-part character (a minor character) in the improvisation; (b) drive the animations of the avatars in the user interface so that they react bodily in ways that is consistent with the affect that they are expressing, for instance by changing posture or facial expressions.

Much research has been done on creating affective virtual characters in interactive systems. Indeed, Picard's work [1] makes great contributions to building affective virtual characters. Also, emotion theories, particularly that of Ortony, et al. [2] (OCC), have been used widely in such research. Egges et al. [3] provided virtual characters with conversational emotional responsiveness. Aylett et al. [4] also focused on the development of affective behaviour planning for their synthetic characters. There is much other work in a similar vein.

Emotion recognition in speech and facial expression has been studied extensively [5, 6]. But very little research work has made an attempt to dig out the affect flavour in human open-ended linguistic textual input in online role-play, although the first interaction system based on natural language textual input, Eliza, was first developed back in 1966. Thus there has been only a limited amount of work directly comparable to our own, especially given our concentration on improvisation and open-ended language. However, Façade [7] included shallow natural language processing for characters' open-ended utterances. But the detection of major emotions, rudeness and value judgements is not mentioned. Zhe and Boucouvalas [8] demonstrated an emotion extraction module embedded in an Internet chatting environment. Unfortunately the emotion detection focuses only on emotional adjectives, and does not address deep issues such as figurative expression of emotion. Also, the concentration purely on first-person emotions is narrow. Our work is thus distinctive in these aspects, including affect detection in metaphorical language and context profiles, and also from first-person and third-person perspectives.

Various characterizations of emotion are used in emotion theories. The OCC model uses emotion labels (anger, etc.) and intensity, while Watson and Tellegen [9] use positivity and negativity of affect as the major dimensions. We have drawn ideas from several such sources. We use an evaluation dimension (negative-positive), affect labels, and intensity. The basic emotion labels (such as 'angry') we use are taken from Ekman [10], while other comparatively complex affect labels (such as 'approving') are taken from the OCC model. There are 25 affect labels used in our system currently. Affect labels plus intensity are used when strong text clues signalling affect are detected, while the evaluation dimension plus intensity is used when only weak text clues are detected. In this paper, although first we briefly summarize our previous implementation in section 2.1 & 2.2, we mainly emphasis our new implementations on metaphorical figurative language processing in section 2.3, and affect interpretation based on context in section 2.4 and user testing evaluation for the AI agent and the overall affect sensing component in section 3. We draw conclusion and identify future work in section 4.

2. THE AFFECT DETECTION PROCESSING

Before any automatic recognition and response components could be built for use in our automated AI actor, a detailed analysis of the language used in e-drama sessions was necessary. A small corpus of sessions was analysed by hand to identify the range of linguistic forms used and to provide insight for the automatic processing. In fact, this analysis is often very difficult and unreliable but it does reveal some important observations. The language used in e-drama is complex and idiosyncratic, e.g. often ungrammatical and full of abbreviations, mis-

spellings, etc. Moreover, the language contains a large number of weak cues to the affect that is being expressed. These cues may be contradictory or they may work together to enable a stronger interpretation of the affective state. In order to build a reliable and robust analyser of affect it is necessary to undertake several diverse forms of analysis and to enable these to work together to build stronger interpretations.

2.1 Pre-processing Modules

The language created in e-drama sessions severely challenges existing language-analysis tools if accurate semantic information is sought, even in the limited domain of restricted affect-detection. Aside from the complications noted above, the language includes slang, use of upper case and special punctuation (such as repeated exclamation marks) for affective emphasis, repetition of letters, syllables or words for emphasis, and open-ended interjective and onomatopoeic elements such as “hm”, “ow” and “grrrr”. To deal with the misspellings, abbreviations, letter repetitions, interjections and onomatopoeia, several types of pre-processing occur before the main aspects of detection of affect. We have reported our work on pre-processing modules to deal with these language phenomena in detail in [11, 25].

2.2 Affect Detection using Rasp, Pattern Matching & WordNet and Responding Regimes

One useful pointer to affect is the use of imperative mood, especially when used without softeners such as ‘please’ or ‘would you’. Strong emotions and/or rude attitudes are often expressed in this case. Expression of the imperative mood in English is surprisingly various and ambiguity-prone. We have used the syntactic output from the Rasp parser [12] and semantic information in the form of the semantic profiles for the 1,000 most frequently used English words [13] to deal with certain types of imperatives.

In an initial stage of our work, affect detection was based purely on textual pattern-matching rules that looked for simple grammatical patterns or templates partially involving specific words or sets of specific alternative words. This continues to be a core aspect of our system but we have now added robust parsing and some semantic analysis, including but going beyond the handling of imperatives discussed above.

A rule-based Java framework called Jess is used to implement the pattern/template-matching rules in the AI agent allowing the system to cope with more general wording. This procedure possesses the robustness and flexibility to accept many ungrammatical fragmented sentences. The rules conjecture the character’s emotions, evaluation dimension (negative or positive), politeness (rude or polite) and what response the automated actor should make. However, it lacks other types of generality and can be fooled when the phrases are suitably embedded as subcomponents of other grammatical structures. In order to go beyond certain such limitations, sentence type information obtained from the Rasp parser has also been adopted in the pattern-matching rules. This information not only helps the AI agent to detect affective states in the user’s input (such as the detection of imperatives), and to decide if the detected affective states should be counted (e.g. affects detected in conditional sentences won’t be valued), but also contributes to proposing appropriate responses.

Additionally, a reasonably good indicator that an inner state is being described is the use of ‘I’, especially in combination with the present or future tense (e.g. ‘I’ll scream’, ‘I hate/like you’, and ‘I need your help’). We especially process ‘the first-person with a present-tense verb’ statements using WordNet.

We have also created responding regimes for the AI character. Most importantly, the AI agent can adjust its response likelihood according to how confident the AI agent is about what it has discerned in the utterance at hand. Especially, in order to make contributions to the improvisation progression, the AI agent also has a global view of the drama improvisation. Briefly, the knowledge base of the AI actor provides scenario’s background knowledge for each human character. The AI agent can raise various scenario-related topics in its role for the human characters according to the detected affective states and topics discussed in the text input by

using the rule-based reasoning based on the knowledge base. Inspection of the transcripts collected in the user testing indicates that the AI actor usefully pushed the improvisation forward on various occasions (see section 3). Details of the work reported in this section can be found in [11, 25].

2.3 Metaphorical Language Understanding in the AI Actor

The metaphorical description of emotional states is common and has been extensively studied [14, 15]. E.g.: "He nearly exploded" and "Joy ran through me," where anger and joy are being viewed in vivid physical terms. Such examples describe emotional states in a relatively explicit if metaphorical way. But affect is also often conveyed more implicitly via metaphor, as in "His room is a cess-pit": affect (such as 'disgust') associated with a source item (cess-pit) gets carried over to the corresponding target item (the room). In other work, we have conducted research on metaphor in general (see, e.g. [16, 17]), and are now applying it to this application, and conversely using the application as a useful source of theoretical inspiration.

In our collected transcripts, metaphorical language has been used extensively to convey emotions and feelings. One category of affective metaphorical expressions that we're interested in is 'Ideas/Emotions as Physical Objects" [16, 17], e.g. "joy ran through me", "my anger returns in a rush", "fear is killing me" etc. In these examples, emotions and feelings have been regarded as external entities. The external entities are often, or usually, physical objects or events. Therefore, affects could be treated as physical objects outside the agent in such examples, which could be active in other ways [16]. Implementation has been carried out to provide the affect detection component the ability to deal with such affect metaphor. We mainly focus on the user input with the following structures: 'a singular common noun subject + present-tense lexical verb phrase' or 'a singular common noun subject + present-tense copular form + -ing form of lexical verb phrase'. WordNet-affect domain (part of WordNet-domain 3.2) [18] has been used in our application. It provides an additional hierarchy of 'affective domain labels', with which the synsets representing affective concepts are further annotated (e.g. 'panic' is interpreted as 'negative-fear -> negative-emotion -> emotion -> affective-state -> mental-state'). Also with the assistance of the syntactic parsing from Rasp, the input "panic drags me down" is interpreted as 'a mental state + an activity + object (me)'. Thus the system regards such expression as affective metaphor belonging to the category of 'affects as entities'.

In daily expressions, food has been used extensively as metaphor for social position, group identity, religion, etc. E.g. food could also be used as a metaphor for national identity. British have been called 'roastbeefs' by the French, while French have been referred to as 'frogs' by the British. In one of the scenarios we used (school bullying), the big bully has called the bullied victim (Lisa) names, such as "u r a pizza", "Lisa has a pizza face" to exaggerate that fact that the victim has acne. Another most commonly used food metaphor is to use food to refer to a specific shape. E.g. body shape could be described as 'banana', 'pear' and 'apple' (http://jsgfood.blogspot.com/2008/02/food-metaphors.html). In our application, "Lisa has a pizza face" could also be interpreted as Lisa has a 'round (shape)' face. Therefore, insults could be conveyed in such food metaphorical expression. We especially focus on the statement of 'second-person/a singular proper noun + present-tense copular form + food term' to extract affect. A special semantic dictionary has been created by providing semantic tags to normal English lexicon. The semantic tags have been created by using Wmatrix [19], which facilitates the user to obtain corpus annotation with semantic and part-of-speech tags to compose dictionary. The semantic dictionary created consists mainly of food terms, animal names, measureable adjectives (such as size) etc with their corresponding semantic tags due to the fact they have the potential to convey affect and feelings.

In our application, Rasp informs the system the user input with the desired structure - 'second-person/a singular proper noun + present-tense copular form + noun phrases' (e.g. "Lisa is a pizza", "u r a hard working man", "u r a peach"). The noun phrases are examined in order to recover the main noun term. Then its corresponding semantic tag is derived from the composed semantic dictionary if it is a food term, or an animal-name etc. E.g. "u r a peach" has been

regarded as "second-person + present-tense copular form + [food-term]". WordNet [20] has been employed in order to get the synset of the food term. If among the synset, the food term has been explained as a certain type of human being, such as 'beauty', 'sweetheart' etc. Then another small slang-semantic dictionary collected in our previous study containing terms for special person types (such as 'freak', 'angle') and their corresponding evaluation values (negative or positive) has been adopted in order to obtain the evaluation values of such synonyms of the food term. If the synonyms are positive (e.g. 'beauty'), then we conclude that the input is an affectionate expression with a food metaphor (e.g. "u r a peach").

However, in most of the cases, WordNet doesn't provide any description of types of human beings when explaining a food term (e.g. 'pizza', 'meat' etc). According to the nature of the scenarios (e.g. bullying) we used, we simply conclude that the input implies insulting with a food metaphor when calling someone food terms ("u r walking meat", "Lisa is a pizza").

Another interesting phenomenon drawing our attention is food as shape metaphor. As mentioned earlier, food is often used as a metaphor to refer to body shapes (e.g. "you have a pear body shape", "Lisa has a garlic nose", "Lisa has a pizza face"). They might indicate literal truth, but most of which are potentially used to indicate very unpleasant truth. Thus they could be regarded as insulting. We extend our semantic dictionary created with the assistance of Wmatrix by adding terms of physiological human body parts, such as face, nose, body etc. For the user's input with a structure of 'second-person/a singular proper noun + have/has + noun phrases' informed by Rasp, the system provides a semantic tag for each word in the object noun phrase. If the semantic tag sequence of the noun phrase indicates that it consists of a food term followed by a physiological term ('pizza face'), the system interprets that the input implies insulting with a food metaphor.

However, examples, such as "you have a banana body shape" and "you are a meat and potatoes man", haven't been used to express insults, but instead the former used to indicate a slim body and the latter to indicate a hearty appetite and robust character. Other examples such as "you are what you eat" could be very challenging theoretically and practically. In order to gain more flexibility and generalization when dealing with metaphorical expressions, we have also used a statistical-based machine learning approach to conduct some experiments on the recognition of the above affect and food metaphors.

2.4 Context-based Affect Detection

Our previous affect detection has been performed solely based on individual turn-taking user input. Thus the context information has been ignored. However, the contextual and character profiles may influence the affect conveyed in the current input. In this section, we are going to discuss cognitive emotion simulation for individual characters and contextual emotion modeling for other characters' influence towards the current speaking character in communication context and our approach developed based on these features to interpret affect from context.

In our study, we previously noticed some linguistic indicators for contextual communication in the recorded transcripts. E.g. one useful indicator is (i) imperatives, which are often used to imply negative or positive responses to the previous speaking characters, such as "shut up", "go on then", "let's do it" and "bring it on". Other useful contextual indicators are (ii) prepositional phrases (e.g. "by who?"), semi-coordinating conjunctions (e.g. "so we are good then"), subordinating conjunctions ("because Lisa is a dog") and coordinating conjunctions ('and', 'or' and 'but'). These indicators are normally used by the current 'speaker' to express further opinions or gain further confirmation from the previous speakers.

In addition, (iii) short phrases for questions are also used frequently in the transcripts to gain further communication based on context, e.g. "where?", "who is Dave" or "what". (iv) Character names are also normally used in the user input to indicate that the current input is intended for particular characters, e.g. "Dave go away", "Mrs Parton, say something", "Dave what has got into you?" etc. Very often, such expressions have been used to imply potential emotional contextual

communication between the current speaking character and the named character. Therefore the current speaking characters may imply at least 'approval' or 'disapproval' towards the opinions/comments provided by the previous named speaking characters. Finally there are also (i) some other well known contextual indicators in Internet relay chat such as 'yeah/yes followed by a sentence ("yeah, we will see", "yeah, we cool Lisa")', "I think so", 'no/nah followed by a sentence', "me too", "exactly", "thanks", "sorry", "grrrr", "hahahaha", etc. Such expressions are normally used to indicate affective responses to the previous input. However, these linguistic indicators act as very limited signals for contextual communication. There are still cases ("ur a batty 2 then okay", "the rest dropped out cuz they didn't want to play with a gay", "I want to talk about it now") that contextual affect analysis fails to be activated to derive affect implied in the user's input. In the work reported here, we intend to deal with such difficulties by activating contextual affect analysis even for input with structures of "subjects + verb phrases + objects". Especially an input with a structure of 'second person + copular form (you are)' tends to convey insulting in our application ("u r a batty 2 then okay", "u r walking meat" etc).

2.4.1 Emotion Modeling using Bayesian Networks

Lopez et al. [26] has suggested in their work that context profiles for affect detection have been referred to social, environmental and personal contexts. In our study, personal context may be regarded as one's own emotion inclination or improvisational mood in communication context. Bayesian networks have been used to simulate such personal emotion context. E.g. in this Bayesian network, we regard the first emotion experienced by a user as A, the second emotion experienced by the same user as B, and the third emotion experienced as C. We believe that one's own emotional states have a chain reaction effect. For example, the previous emotional status may influence later emotional experience. We have made attempts to embody such chain effects into emotion modeling for personal context. We assume that the second emotional state B, in any combination is dependent on the first emotional state A. Further, we assume that the third emotional state C, is dependent on both the first and second emotional states A and B. In our application, if we only consider two most recent emotional states the user experiences as the most related relevant context based on Relevance theory [21, 22], then we may predict what the most probable emotion the user is the most likely to experience in the next turn-taking using a Bayesian network.

A Bayesian network employs a probabilistic graphical model to represent causality relationship and conditional (in)dependencies between domain variables. It allows combining prior knowledge about (in)dependencies among variables with observed training data via a directed acyclic graph. It has a set of directed arcs linking pairs of nodes: an arc from a node X to a node Y means that X (parent emotion) has a direct influence on Y (successive emotion). Such causal modeling between variables reflects the chain effect of emotional experience. It uses the conditional probability (P[B|A], P[C|A,B]) to reflect such influence between prior emotional experiences to successive emotional expression. The following network topology has been used to model personal contextual emotional profiles in our application.

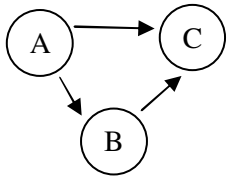

FIGURE 1: An Emotion Network

In Figure 1, conditional probabilities are needed to be calculated for the emotional state C given any combination of the emotional states A and B. Theoretically, emotional states A and B could be any combination of potential emotional states. Similarly, since there could be several

emotional states considered as successive emotional state C, we have considered a conditional probability for each potential successive emotional state. In our application, we have mainly considered the following 10 most frequently used emotional states for the simulation of the improvisational mood for a particular character in the Bayesian network: 'neural', 'happy', 'approval', 'grateful', 'caring', 'disapproval', 'sad', 'scared', 'insulting', and 'angry'. Any combination of the above emotional states could be used as prior emotional experience of the user. Altogether the overall combinations for the two prior emotions are counted as 100 (10 * 10). Also each conditional probability of each emotional state in the above given two prior emotional experiences (such as P[happy| A,B], P[approval| A,B] etc) will be calculated as the confidence for later selection. Then the emotional state with the highest conditional probability, P[C|A,B], will be chosen as the most probable emotional experience the user may express in his/her very next turn-taking. In this way, we model contextual emotional chain effect for an individual character to benefit our contextual affect detection.

An advantage of using Bayesian networks for emotion simulation and modeling is that it is not necessary to gather training data from other sessions of the same scenarios to train the system at the beginning to allow future prediction. We can simply use the emotional states experienced by a particular character throughout the improvisation as the prior input emotions to the Bayesian network so that our system may learn about this user's emotional trend and mood gradually without any constrains set by the training data or scenario related information.

Moreover we also take a frequency approach to determine the conditional probabilities. When an affect has been detected from the user's input, we increment a counter for that expressed emotion given the two prior implied emotional states. An example conditional probability table has been shown in Table 1.

		Probability of the predicted emotional state C being:			
Emotion A	Emotion B	Happy	Approval	...	Angry
Happy	Neutral	P_{00}	P_{01}	...	P_{09}
Neutral	Angry	P_{10}	P_{11}	...	P_{19}
Disapproval	Disapproval	P_{20}	P_{21}	...	P_{29}
Angry	Angry	P_{30}	P_{31}	...	P_{39}

TABLE 1: Conditional Probability Table for Emotions Expressed

When making a prediction for an emotion state mostly likely to be shown in the very next input for one particular character, the two prior emotional states are used to determine which row to consider in the conditional probability matrix, and select the column with the highest conditional probability as the final output. Example conditional probability calculations are shown in the following formulas, where N represents the total number of emotions shown so far by this particular character and N with a subscript indicates the number of a particular emotion shown given previously expressed emotions. E.g., $N_{happy\_neutral\_happy}$ indicates the occurrences that two prior emotions A and B are respectively happy & neutral and the following emotional state C is happy.

$P(A = happy) = N_{happy}/N$
$P(B = neutral) = N_{neutral}/N$
$P(B = neutral| A = happy) = N_{neutal\_happy}/N$
$P(C = happy| A = happy, B = neutral) = N_{happy\_neutral\_happy}/N_{AB}$

As we mentioned earlier, the probabilities are not necessarily to be produced by any training data and stored in advance. The frequencies are sufficient to use to calculate probabilities when required. In our case, we store the frequencies of emotion combinations in a 100 * 10 ((A*B) * C) matrix dynamically.

In our application, one of the scenarios has been used for user testing is Homophobic bullying. We briefly introduce this scenario in the following since example transcripts have been taken from this scenario for the discussion of the contextual affect detection implementation reported here. The character Dean (16 years old), captain of the football team, is confused about his sexuality. He has ended a relationship with a girlfriend because he thinks he may be gay and has told her this in confidence. Tiffany (ex-girlfriend) has told the whole school and now Dean is being bullied and concerned that his team mates on the football team will react badly. He thinks he may have to leave the team. The other characters are; Tiffany who is the ring leader of the bullying, and wants Dean to leave the football team, Rob (Dean's younger brother) who wants Dean to say he is not gay to stop the bullying, Lea (Dean's older sister) who wants Dean to be proud of who he is and ignore the bullying, and Mr Dhanda (PE Teacher) who needs to confront Tiffany and stop the bullying.

Suppose we have the following sequence of example interaction extracted from the recorded transcripts for the Tiffany character in this scenario. Based on the affect detection purely from the analysis of each individual input, we obtained the emotional states implied in the first three inputs from Tiffany as the following: 'angry, angry, and angry'.

Tiffany: Dean, U R DISGUSTING!! u shuld leav da football team. [angry]
...
Tiffany: shut up man lea [angry]
...
Tiffany: u get out of here. [angry]
...
Tiffany: ur a batty 2 then okay [neutral] -> [angry]

Also we have derived 'neutral' for the very last input without any contextual inference. Since the input has a structure of "second person + copular form", as discussed earlier which is very often used to convey insulting or compliment in our application, the context-based affect analysis will be activated to adjust/predict the affect conveyed in the last input from the above example transcript. This emotional sequence implied by Tiffany ('angry, angry, and angry') will be used to 'train' the contextual emotional simulation and construct the Bayesian probability matrix, which will be used to predict the most probable emotion implied in Tiffany's very last input. In this example, we need to calculate the conditional probability of P[C| angry, angry, angry] for each potential emotional state C. Finally the emotional state 'angry' has achieved the highest probability result and been predicted as the most probable emotion implied in the input "ur a batty 2 then okay". Thus we adjust the emotional state for the very last input from 'neutral' to 'angry'.

Therefore in this way, we can produce emotion modeling for each individual character within the same and across scenarios. However, other contextual profiles (such as other characters' emotional profiles and discussion topics) are still needed to further justify the affect detected using the above discussed Bayesian network approach. In the following section, we introduce how social emotional contextual profiles are used to model emotional influence from other characters to the current speaking character during the improvisation.

2.4.2 Emotional Social Context Modeling using Unsupervised Neural Networks

The simulation of one's own emotional context and improvisational mood is important, but the modeling of other characters' emotional influence to the current speaking character is also crucial for the accurate interpretation of the affect implied in the current input. For example, the emotional context contributed by other participants, e.g. friend or enemy characters, may (dramatically) affect the speaking character's emotional expression in the next turn-taking in our application. Moreover, a discussion topic or an improvisation is composed of the mixture of such emotional sub-contexts. They take the overall forms of being positive or negative and have been acted as the most relevant emotional social context to the current speaking character. If such social positive/negative most relevant context could be recognized during the improvisation, it is very helpful to justify the affective states detected from personal context modeling using the

above discussed Bayesian approach. In order to recognize the positive/negative trend in the most related sub-context contributed by (part of) participants, an unsupervised neural network learning algorithm has been employed. I.e. Adaptive Resonance Theory-1 (ART1) has been used in our application to derive general emotional implication (positive/negative/neutral) for the most recent interaction context.

Generally, ART is a collection of models for unsupervised learning, which deals with object identification and recognition generally as a result of the interaction of 'top-down' observer expectations with 'bottom-up' sensory information. The 'top-down' template or prototype will be used to compare with the actual features of an object as detected by the senses to produce categorizations for the observed objects. ART-1 in particular has been used to resolve stability and plasticity dilemma, i.e. the ability to maintain previously learned knowledge ('stability') while still being capable of learning new information ('plasticity'). Although it mainly accepts binary input vectors, this is sufficient enough in our application currently. In our application, it would be beneficial that the positive/negative context prediction modeling is capable of both retaining previously learned information (e.g. the sensing of positive or negative context in a particular scenario) and in the meantime integrating newly discovered knowledge (e.g. the sensing of such context across different scenarios). Such capability may allow the emotional social context modeling to perform across scenarios. Also, the ART-1 algorithm has an advanced ability to create a new cluster when required with the assistance of a vigilance parameter. It may help to determine when to cluster a feature vector to a 'close' cluster or when a new cluster is needed to accommodate this feature vector.

In our application, we use the evaluation values (positive and negative) and neutralization of the most recent several turn-taking as the input to ART-1. In detail, for each user input, we convert its emotional implication into pure positive or negative and use three binary values (0 or 1) to represent the three emotional implications: neutral, positive and negative. For example, for the input from Arnold in the Crohn's disease scenario (another scenario used in our application), "dont boss me about wife [angry]" when the wife character, Janet, was too pushy towards the husband character, Arnold. We have used '0 (neutral), 0 (positive), and 1 (negative)' to indicate the emotional inclination ('angry' -> 'negative') in the user input. Another example transcript taken from the Homophobic bullying scenario is listed in the following.

1. Tiffany Tanktop: sorry, all io could hear was I'M A BIG GAY [insulting/angry]
2. Mr. Dhanda: TIFFANY I WILL....GET YOU EXPENDED IF YOU DONT FOLLOW MY ORDERS! YOU HOMO-FOBIC [angry]
3. Rob Hfuhruhurr: tiffany is wierd lol y she spreadn rumors etc???? [disapproval]
4. Tiffany Tanktop: there not rumours...its the truth [disapproval]
5. Tiffany Tanktop: GGGGAAAYYYYY! [insulting/angry]
6. Mr. Dhanda: TIFFANY STOP IT NOW!!! [angry]
7. Mr. Dhanda: ILL BANG YOU [angry]
8. Rob Hfuhruhurr: god leav hm alone!!! [angry]
9. Tiffany Tanktop: ONCE A BATTY ALWAYS A BATTY [neutral] -> [angry]

For the very last input from Tiffany, we can only interpret 'neutral' based on the analysis of the input itself without using any contextual inference although emotional states have been derived for all the other input based on the analysis the input themselves. In order to further derive/justify the affect conveyed in the very last 'neutral' input although there is no any linguistic indicator for contextual communication existing, we resort to the prediction of the general emotional trend using the most related interaction context contributed by several participant characters. Since normally in one session, up to 5 characters are involved in the improvisation as mentioned previously, except for the last input, we have taken the previous last four inputs to the current last input as the most related context for prediction of the positive/negative inclination in the social context. Thus we have taken the input from Rob (8th input), Mr Dhanda (7th and 6th input), and Tiffany (5th input) for consideration and prediction. Since Tiffany implies 'angry' (binary value combination for neutral, positive and negative: 001) by saying "GGGGAAAYYYYY!", Mr Dhanda

also indicating 'angry' (001) in both of his input: "TIFFANY STOP IT NOW!!!" and "ILL BANG YOU", followed by another 'angry' (001) input from Rob "god leav hm alone!!!", we have used the following feature vector to represent this most related discussion context: '001 001 001 001 (angry, angry, angry and angry)'. This feature vector is used as the input to the ART-1 unsupervised learning algorithm to determine its category belongingness. In a similar way, we can gather a set of such feature vectors from the same and across scenarios. ART-1 will classify these feature vectors into different groups based on their similarities and differences shown in the vectors, which sometimes may not be apparent at the beginning.

Briefly, we begin the ART-1 algorithm with a set of unclustered emotional context feature vectors and some number of clusters. For each emotional feature vector, ART-1 makes attempts to find the cluster to which it's closest. A similarity test calculates how close the feature vector to the cluster vector. The higher the value, the closer the vector is to the cluster. If a feature vector is sufficiently close to a cluster, we then test for vigilance acceptability, which is the final determiner for whether the feature vector should be added to the particular cluster. If a feature vector also passes the vigilance test, then we assign it to and update that particular cluster with the features of the new addition. If a feature vector fails the similarity test or vigilance test for all the available clusters, then a new cluster is created for this feature vector. When new clusters are created, some feature vectors may drop out of a cluster and into another based on new feature vectors being added and adjusting the cluster vector. Thus ART-1 will start the process again by checking through all the available feature vectors. If no feature vector needs to change its cluster, the process is complete. In our application, we can gradually feed emotional context feature vectors to ART-1, which will not only remain the previous classification of positive or negative context in a particular scenario, but also indefinitely integrate new positive/negative context extracted from other interaction across scenarios. Suppose we have the following emotional contexts contributed by the Crohn's disease scenario and classified previously by ART-1 into three categories:

Class 0 contains:

0 [1 0 0 0 0 1 0 0 1 0 0 1] negative1 (Neutral, sad, disapproving and sad)
1 [1 0 0 0 1 0 0 0 1 0 0 1] negative2 (Neutral, approving, disapproving and angry)
2 [1 0 0 0 0 1 0 0 1 1 0 0] negative3 (neutral, disapproving, disapproving and neutral)
3 [0 0 1 0 1 0 0 0 1 0 0 1] negative4 (angry, approving, disapproving, and angry)
5 [1 0 0 0 0 1 0 0 1 0 1 0] negative6 (neutral, angry, angry and approving)

Class 1 contains:

4 [0 0 1 0 0 1 1 0 0 1 0 0] negative5 (angry, angry, neutral and neutral)
8 [1 0 0 0 1 0 1 0 0 0 0 1] positive3 (neutral, caring, neutral and disapproval)
9 [1 0 0 1 0 0 1 0 0 1 0 0] neutral1 (neutral, neutral, neutral and neutral)

Class 2 contains:

6 [0 1 0 0 1 0 0 1 0 1 0 0] positive1 (happy, happy, happy and neutral)
7 [1 0 0 0 1 0 0 1 0 0 1 0] positive2 (neutral, caring, approving and happy)
10 [0 1 0 0 1 0 1 0 0 0 1 0] positive4 (approval, grateful, neutral and approval)

Since ART-1 is not aware which label it should use to mark the above each category although it classifies the emotional feature vectors based on their similarities and differences and achieves the above classification, a simple algorithm will make attempts to assign labels (positive/negative/neutral context) to the above classification based on the majority vote of the evaluation values of all the emotional states shown in each feature vector in each category. For example, Class 0 has assigned 4 emotional feature vectors and most of the emotional states in all the feature vectors in this category are 'negative', therefore it is labelled as 'negative context'. Similarly Class 1 is recognised as 'neutral context' with Class 2 identified as 'positive context'. If we add the above example emotional context from the Homophobic bullying scenario as a new

feature vector, '001 001 001 001' (angry, angry, angry and angry), to the algorithm, we have Class 0 updated to accommodate the newly arrived feature vector as output. Thus the new feature vector is 'classified' as 'negative context'. Therefore, the last input from Tiffany ("ONCE A BATTY ALWAYS A BATTY") is more likely to contain 'negative' implication rather than 'neutral' based on the consideration of its most relevant emotional context.

In our application, the context-based affect analysis normally activates the personal context modeling using the Bayesian networks first and then follows the emotional social context modeling using ART-1 to justify or further derive the affect conveyed in the current input. For example, in the above Homophobic bullying example transcript, the emotional context of Tiffany is retrieved as 'angry (1st input), disapproval (4th input) and angry (5th input)'. Thus we use the Bayesian network first to predict the most likely affective state conveyed in Tiffany's very last input. The emotional state 'angry' has achieved the highest probability and been regarded as the affect mostly likely implied in the input ("ONCE A BATTY ALWAYS A BATTY"). Since the most relevant discussion context contributed by the 5th – 8th input is also sensed as being 'negative' using the ART-1 approach discussed above, we conclude that the very last input from Tiffany is more likely to be 'angry' with a strong intensity indicated by the capitalization. Thus we adjust the affect implied in the very last input from 'neutral' to 'angry'.

In this way, we can predict the next most probable emotional state based on a character's previous emotional implications during the improvisation using the Bayesian networks and detect the 'positive or negative' emotional implication of the most related discussion context using unsupervised learning. The integration of both discussed approaches has great potential to derive affect in communication context which is closer to the user's real emotional experience. Another advantage of our implementation is that it has the potential to perform contextual affect sensing across different scenarios.

At the test stage, our affect detection component integrated with the AI agent detects affect for each user input solely based on the analysis of individual turn-taking input itself as usual. The above algorithms for context-based affect sensing will be activated when the affect detection component recognizes 'neutral' from the current input during the emotionally charged proper improvisation and the input also containing statement structures.

In this way, by considering the potential improvisational mood one character was in and recent social emotional profiles of other characters, our affect detection component has been able to inference emotion based on context to adjust the affect interpreted by the analysis based on individual turn-taking user input. After the description of various affect processing components, the overall affect detection model is shown in Figure 2.

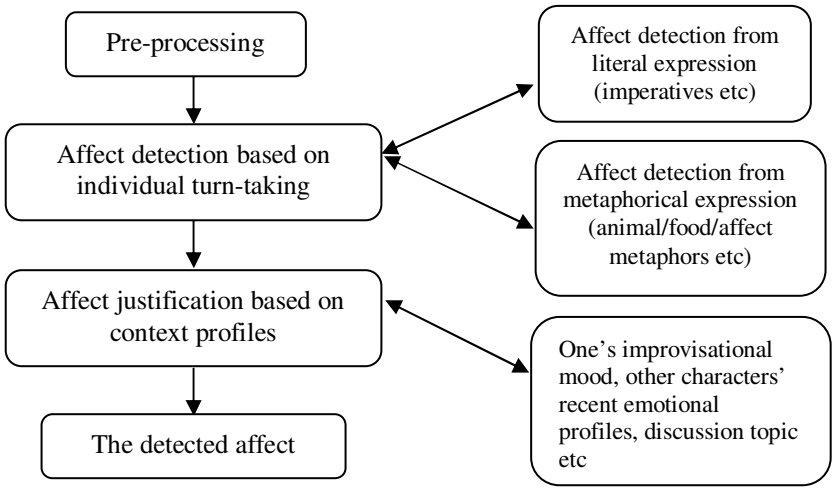

FIGURE 2: The Affect Detection Model

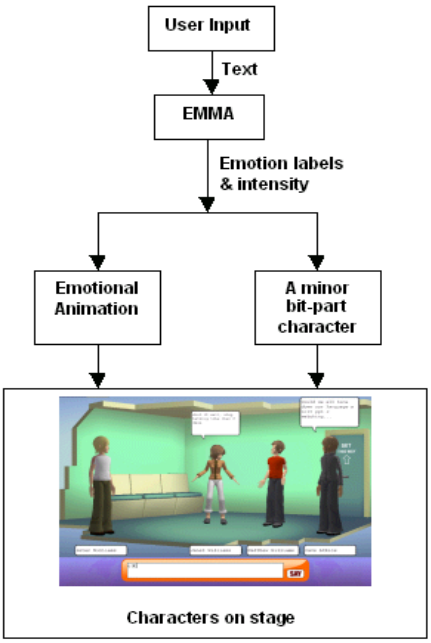

FIGURE 3: Affect Detection and the Control of Characters

The detected affective states from users' open-ended text input play an important role in producing emotional animation of human players' avatars. The emotional animation mainly includes emotional gesture and social attention (such as eye gazing). The expressive animation engine, Demeanour [23], makes it possible for our characters to express the affective states detected by the AI actor, EMMA. When it detects an affective state in a user's text input, this is passed to the Demeanour system attached to this user's character and a suitable emotional animation is produced. The Demeanour system has also used character profiles, particularly including personality traits and relationships with other characters, to provide expressive animation for other avatars when the 'speaking' avatar experiences affect. How the detected affective states inform the animation engine and control the AI agent is illustrated in Figure 3.

3. USER TESTING OF THE IMPROVISATIONAL AI ACTOR

We conducted an intensive user test with 160 secondary school students, in order to try out and refine a testing methodology. The aim of the testing was primarily to measure the extent to which having the AI agent as opposed to a person play a character affects users' level of enjoyment, sense of engagement, etc. We concealed the fact that the AI-controlled agent was involved in some sessions in order to have a fair test of the difference that is made. We obtained surprisingly good results. Having a minor bit-part character called "Dave" played by the AI agent as opposed to a person made no statistically significant difference to measures of user engagement and enjoyment, or indeed to user perceptions of the worth of the contributions made by the character "Dave". Users did comment in debriefing sessions on some utterances of Dave's, so it was not that there was a lack of effect simply because users did not notice Dave at all. Furthermore, it surprised us that few users appeared to realize that sometimes Dave was computer-controlled. We stress, however, that it is not an aim of our work to ensure that human actors do not realize this.

Inspection of the transcripts collected indicates that the AI agent usefully pushed the improvisation forward on various occasions. Figure 4 shows an example about how the AI actor contributed to the drama improvisation in Crohn's disease scenario. Briefly, in Crohn's disease scenario, Peter has had Crohn's disease since the age of 15. Crohn's disease attacks the wall of the intestines and makes it very difficult to digest food properly. The character has the option to undergo surgery (ileostomy) which will have a major impact on his life. The task of the role-play is to discuss the pros and cons with friends and family and decide whether he should have the operation. The other characters are; Mum, who wants Peter to have the operation, Matthew (older brother) who is against the operation, Dad who is not able to face the situation, and David (the best friend) who mediates the discussion. In the example transcript shown in Figure 4, Dave was played by the AI actor, which successfully led the improvisation on the desirable track. In another scenario, the Homophobic bullying, used for the testing, Mr. Dhanda was sometimes played by EMMA and example transcripts are also shown that the AI actor has helped to push the improvisation forward.

We have conducted an initial evaluation of the quality of the AI agent's determinations about emotion during these testing sessions, by comparing the AI agent's determinations during one of the Crohn's disease improvisations with emotion labels later assigned offline by two members of our team (not involved in the development of the AI agent's algorithms). We used the kappa statistic of Carletta [24]. It is a measure of the pairwise agreement among a set of coders making category judgements, correcting for expected chance agreement. The statistic, K, is calculated as $K= (P(A)-P(E))/(1-P(E))$ where P(A) is the proportion of times two coders agree and P(E) is the proportion of times we would expect them to agree if they categorized randomly. A value of at least 0.6 – 0.8 is generally required by researchers looking for good inter-annotator agreement. We calculated K for each pair among the three labellers (EMMA and two humans). The inter-human K was only 0.35, and so it is not surprising that the EMMA/human values were only 0.34 and 0.32. Although they are not ideal, at least these results give grounds for hope that our affect detection with further refinement can come near the rather low human/human level of agreement. Moreover, the overall accuracy rate of affect interpretation based on context in our current analysis achieves 68% via the comparison of the annotation of part of the recorded transcripts between human annotators and the AI agent in the Crohn's disease scenario.

From the inspection of the evaluation results, although context affect detection is provided, there are still some cases: when the two human judges both believed that user inputs carried negative affective states (such as angry, threatening, disapproval etc), EMMA regarded them as neutral. One most obvious reason is that some of the previous pipeline processing (such as dealing with mis-spelling, acronyms etc, and syntactic processing from Rasp etc) failed to recover the standard user input or recognize the complex structure of the input which led to less interesting and less emotional context for some of the characters and may affect the performance of contextual affect sensing. We also aim to extend the evaluation of the context-based affect detection using transcripts from other scenarios. Also, since the test transcripts contained a very

small number of metaphorical language phenomena comparatively, we intend to use other resources (e.g. Wallstreet Journal and other metaphorical databases (e.g. ATT-Meta)) to further evaluate the new development on metaphorical affect sensing.

Example transcripts - Crohn's disease scenario
Dave - the AI character

Peter Williams: I'M THE ILL ONE
Dave Atkins: Arnold, Peter needs ur support and u can't just ignore it.
Janet Williams: nobody cares peter
Arnold Williams: we know peter now stop talking about it
Peter Williams: yes i need your support
Dave Atkins: Hey guys, lets all calm down a bit.
Arnold Williams: we will discuss it later
Janet Williams: well you have got mine
Peter Williams: help me daddy
Arnold Williams: not now son

FIGURE 4: Part of an Example Transcript Produced in the User Testing

The preliminary results from statistical analysis of the collected post questionnaires also indicate that when the AI actor is involved in the improvisation, users' abilities to concentrate on the improvisation are somewhat higher in Crohn's disease scenario than Homophobic bullying scenario. When the AI actor is not involved in the improvisation, users' abilities to concentrate on the improvisation are a lot higher in Homophobic bullying than Crohn's disease. This seems very interesting, as it seems to be showing that the AI actor can make a real positive difference to an aspect of user engagement when the improvisation is comparatively uninteresting.

4. CONCLUSIONS

Our work makes a contribution to the issue of what types of automation should be included in interactive narrative environments, and as part of that the issue of what types of affect should be detected (by directors, etc.) and how. Moreover, our work also makes a contribution to the development of automatic understanding of human language and emotion. Our contextual affect sensing shows initial directions for emotion modeling in personal and social context across scenarios. Future work could include the equipment of the AI agent with the ability of performing autonomous learning through metaphorical expressions.

5. REFERENCES

1. R.W. Picard. *"Affective Computing"*. The MIT Press. Cambridge MA. 2000

2. A. Ortony, G.L. Clore & A. Collins. *"The Cognitive Structure of Emotions"*. Cambridge U. Press. 1998

3. A. Egges, S. Kshirsagar & N. Magnenat-Thalmann. *"A Model for Personality and Emotion Simulation"*, In Proceedings of Knowledge-Based Intelligent Information & Engineering Systems (KES2003), Lecture Notes in AI. Springer-Verlag: Berlin. 2003

4. R.S. Aylett, J. Dias and A. Paiva. *"An affectively-driven planner for synthetic characters"*. In Proceedings of ICAPS. 2006

5. Nogueiras et al. *"Speech emotion recognition using hidden Markov models"*. In Proceedings of Eurospeech 2001, Denmark. 2001

6. M. Pantic, A. Pentland, A. Nijholt and T. Huang. *"Human Computing and Machine Understanding of Human Behavior: A Survey"*. In Proc. Int'l Conf. Multimodal Interfaces, pp. 239-248. 2006

7. M. Mateas. Ph.D. Thesis. "*Interactive Drama, Art and Artificial Intelligence"*. School of Computer Science, Carnegie Mellon University. 2002

8. X. Zhe & A.C. Boucouvalas. *"Text-to-Emotion Engine for Real Time Internet Communication".* In Proceedings of International Symposium on Communication Systems, Networks and DSPs, Staffordshire University, UK, pp 164-168. 2002

9. D. Watson & A. Tellegen. *"Toward a Consensual Structure of Mood".* Psychological Bulletin, 98, 219-235. 1985

10. P. Ekman. *"An Argument for Basic Emotions"*. In Cognition and Emotion, 6, 169-200. 1992

11. L. Zhang, J.A. Barnden, R.J. Hendley & A.M. Wallington. *"Developments in Affect Detection in E-drama".* In Proceedings of EACL 2006, 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006, Trento, Italy. pp. 203-206. 2006

12. E. Briscoe and J. Carroll. *"Robust Accurate Statistical Annotation of General Text"*. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Gran Canaria. 1499-1504. 2002

13. D.R. Heise. *"Semantic Differential Profiles for 1,000 Most Frequent English Words".* Psychological Monographs. 70 8:(Whole 601). 1965

14. S. Fussell & M. Moss. *"Figurative Language in Descriptions of Emotional States".* In S. R. Fussell and R. J. Kreuz (Eds.), Social and cognitive approaches to interpersonal communication. Lawrence Erlbaum. 1998

15. Z. Kövecses. *"Are There Any Emotion-Specific Metaphors?"* In Speaking of Emotions: Conceptualization and Expression. Athanasiadou, A. and Tabakowska, E. (eds.), Berlin and New York: Mouton de Gruyter, 127-151. 1998

16. J. Barnden, S. Glasbey, M. Lee & A. Wallington. *"Varieties and Directions of Inter-Domain Influence in Metaphor"*. Metaphor and Symbol, 19(1), 1-30. 2004

17. J.A. Barnden. *"Metaphor, Semantic Preferences and Context-sensitivity"*. Invited chapter for a Festschrifft volume. Kluwer. 2006

18. C. Strapparava and A. Valitutti. *"WordNet-Affect: An Affective Extension of WordNet"*, In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, 1083-1086. 2004

19. P. Rayson. *"Matrix: A statistical method and software tool for linguistic analysis through corpus comparison".* Ph.D. thesis, Lancaster University. 2003

20. C. Fellbaum. *"WordNet, an Electronic Lexical Database".* The MIT press. 1998

21. D. Sperber & D. Wilson. *"Relevance: Communication and cognition (2nd ed.)"*. Oxford, UK: Blackwell. 1995.

22. D. Wilson & D. Sperber. *"Relevance Theory"*. In G.Ward & L. Horn (Eds.), Handbook of Pragmatics (pp. 607–632). Oxford, UK: Blackwell. 2003.

23. M. Gillies, I.B. Crabtree and D. Ballin. *"Individuality and Contextual Variation of Character Behaviour for Interactive Narrative"*. In Proceedings of the AISB Workshop on Narrative AI and Games. 2006

24. J. Carletta. *"Assessing Agreement on Classification Tasks: The Kappa statistic."* Computational Linguistics, 22 (2), pp.249-254. 1996

25. L. Zhang, M. Gillies, K. Dhaliwal, A. Gower, D. Robertson & B. Crabtree. *"E-drama: Facilitating Online Role-play using an AI Actor and Emotionally Expressive Characters".* International Journal of Artificial Intelligence in Education. Vol 19(1), pp.5-38. 2009

26. J.M. Lopez, R. Gil, R., Garcia, I. Cearreta and N. Garay. *"Towards an Ontology for Describing Emotions".* In WSKS '08 Proceedings of the 1st world summit on The Knowledge Society: Emerging Technologies and Information Systems for the Knowledge Society. 2008